STATE OF CALIFORNIA
DMV
DEPARTMENT OF MOTOR VEHICLES

THE CALIFORNIA DRIVER PERFORMANCE EVALUATION PROJECT: AN EVALUATION OF THE CURRENT DRIVER LICENSING ROAD TEST

By
Nancy Clarke Shumaker

August 1994

Research and Development Section
Division of Program and Policy Administration
California Department of Motor Vehicles

PREFACE

This evaluation is "stage 1" of DMV's master plan to develop a new class C road test. It is being disseminated to interested parties as an internal Research and Development Section document rather than an official publication of the State of California. The findings and opinions contained in the report are those of the author and may not represent the views and policy perspective of the State of California.

ACKNOWLEDGMENTS

The author wishes to acknowledge the individuals who have contributed to this project. The guidance offered by Raymond C. Peck, Chief, Research and Development Section, in the statistical analyses and editing was particularly helpful. Rickey Williams and Robert Hagge, Research Managers, both provided direction for the project. Sharon Seavers, Manager III, acted as liaison between headquarters and the study field offices. Debbie McKenzie, Staff Services Analyst, compiled the final report and provided consistency in the production of tables and drafts.

TABLE OF CONTENTS

PAGE

PREFACE i

ACKNOWLEDGEMENTS.................... ii

INTRODUCTION 1
- Michigan DMP 2
- ADOPT 2
- USC SPT 2
- Michigan DP Variant 3
- California DMV Tests 3

METHODS 3
- Research Design 3
- Study Particulars 5
- Statistical Procedures 5

RESULTS 6
- Sample Description 6
- Fail Rates 7
  - Total sample 7
  - Front-seat examiner tests 8
- Test Result Frequencies and Percentages 9
- Results by Route 10
- Reliability Estimates on Test Means and Passes vs. Failure Rates 11
- Differences Between Offices on Test Scores 13
- Significance of Differences Between Routes and Raters on Test Score 15

DISCUSSION/CONCLUSIONS 16

REFERENCES 17

LIST OF TABLES

NUMBER

1 Current Class C Demographics Percentages 6

2 Current Class C Results for Frequencies, Fail Rate
Average Number Wrong, and Mean Score by Field Office 8

TABLE OF CONTENTS (continued)

LIST OF TABLES (continued)

NUMBER PAGE

3 Current Class C Results for Fail Rate for Front Seat Examiner on First Test 9

4 Current Class C Results for Frequencies and Fail Rate 10

5 Current Class C Results by Route for Fail Rate (Including DQs) 10

6 Current Class C Results by Route for Fail Rate (Excluding DQs) 11

7 Current Class C Reliabilities 11

8 Significance Tests of Differences in Test Reliability by Office (*z*-Transformed Reliability Coefficients) 13

9 Current Class C Correlations Using Pass/Fail as Variable (Excluding DQs) 13

10 Current Class C ANOVA Results of Differences in Mean Score for Front-Seat Examiner on First Test 14

11 Fail Rate Differences on First Test for Front-Seat Examiners (Including DQs) 14

12 Current Class C ANOVA Results for Front-Seat Examiner on First Test 15

13 Fail Rate Differences Between LREs Within Office (Including DQs) 16

INTRODUCTION

The California DMV is currently involved in a comprehensive effort to increase the competency level of the California driving population. One of these efforts involves the development of a new class C (passenger vehicle) drive test. The present report is designed to provide data on the reliability and psychometric properties of the current class C road test in order to provide a baseline comparison for the new drive test. This evaluation of the current test represents "Stage 1" in a multi-phase test development master plan (Williams & Shumaker, 1994).

The above project evolved out of a larger plan to enhance driver competency through improvements in all components of the driver licensing process. The initial phase of the plan involved the commissioning of a consultant to perform a needs and requirements analysis of the California driver licensing process (McKnight & Stewart, 1990). This, in turn, led to the convening of a "Conference on Driver Competency," which brought together nationally and internationally recognized experts on driver licensing and driving behavior (California Department of Motor Vehicles, 1990). The components of California's total driver competency-enhancement plan are summarized in the epilogue to the above proceedings, including the rationale underlying the need for a more reliable and stringent road test.

Test reliability refers to the ability of a test to produce consistent scores or ratings of an individual's level on the performance dimensions or traits reflected in the test. In the case of road tests, there are two sources of potential error or unreliability: interrater reliability and sampling reliability. Interrater reliability is the extent to which LREs assign similar score profiles in rating the same drivers and test behaviors. Difference between raters is obviously not a desirable property since a subject's test score becomes dependent on the particular LRE giving the test. If interrater reliability is very low, the test, in essence, becomes more a measure of differences in LREs than differences between drivers. Sampling reliability is concerned with the adequacy of test length and content in producing a consistent test score. If sampling reliability is high, the applicants should receive similar scores over different test routes or repeated testing on the same route. It is not desirable for sampling reliability to be low since the score an applicant receives is subject to large sampling error and becomes heavily dependent on the specific route or office. The net reliability of the road test is a joint function of the above error sources: rater and sampling. The net reliability will, therefore, be lower than the lowest of these two components.

Another psychometric property of interest is test difficulty and examiner differences in the average score assigned, and pass fail rates. It is possible for interrater reliabilities to be high but for one examiner to score subjects systematically lower or higher than another examiner. This property of the road test was assessed by analyzing differences in average test scores and failure rates by examiner, office and route.

The current test evaluation will not address the issue of test validity per se. Validity is concerned with whether or not a test measures what it purports to measure. Although a reliable test does not guarantee validity, reliability is a necessary condition for validity. An unreliable test cannot be valid.

The following presents a very brief overview of the literature on the reliability of various road tests as contained in Peck (in press).

Michigan DPM

A comprehensive study, the Michigan Driver Performance Measure - DPM, was conducted in 1975 (Forbes et al.). The DPM required 45-60 minutes to complete and focused on search, direction, and speed control abilities rather than placing the usual emphasis on vehicle maneuvering. Drivers earned a total score ("global pattern score") and individual pattern scores ("element scores") for particular driving behaviors.

Interrater reliability was above .90 for global scores and individual element scores. Test-retest reliability was .89 for global scores and .96 for element scores. One must take into account, however, that these high reliabilities are inflated by the pooling of scores of 2 raters and 3 routes and by the application of the Spearman-Brown formula to adjust the reliabilities for differences in test length between the three routes.

Vanosdall et al. (1977) developed a drive test for the Michigan DMV based on the Michigan DPM, but one which halved the time to administer and required only one examiner. Subjects were divided into two groups. One was given the revised DPM, and the other received the standard Michigan drive test. Test-retest (different raters) reliabilities were low, ranging from .40 for the standard test to .59 for the revised DPM test.

ADOPT

In 1981, McPherson and McKnight published a study on a test (ADOPT) they developed for DMV administration in Oklahoma. It was a short test - only ten minutes long. Like the DPM, the ADOPT test used a priori observation sites. Vehicle control, vehicle maneuvering, and interaction with highway traffic were among the 51 behaviors tested.

Interrater reliabilities were .84 for the total test, .83 for skills subscore, and .74 for the practices subscore. Alternate route sampling reliabilities for the two subscores were .76 and .48, respectively.

USC SPT

Jones (1978) developed and evaluated a road test designed to measure skills previously shown to be related to safe driving. The University of Southern California Safe

Performance test was 30 minutes long and required two raters. Like the above tests, the SPT rated drivers at specific points along the drive route. Same-day test-retest administrations resulted in reliabilities of .80 for both a novice and an experienced group of drivers. Because of the 30-minute test length and the requirement of two examiners, the SPT cannot be considered to be an operationally viable test for routine use by DMV agencies (Peck, in press).

Michigan DP Variant

Engel, Paskaruk, and Green (1979) used the Michigan DPM to create a drive test of 41 behaviors. These behaviors were scored either satisfactory or non-satisfactory on speed, search, and direction. Unlike the Michigan DPM, examiners scored subjects along the entire route rather than at specific points on the route. Interrater reliability was reported at .74.

California DMV Tests

In 1978, Ratz created new drive tests for evaluation purposes by lengthening the usual DMV test (experimental test), adding parallel parking (experimental skill test), and increasing the number of points needed to pass. The study found the interrater reliabilities to be .69 for the then current test, .75 for the experimental test, and .78 for the experimental skill test.

Like most state DMV tests, the current California road test is similar to the approach developed and validated by McGlade (1963). Under this approach, scoring is not based on specific observation points or a fixed number of observed behaviors. Instead, the LRE observes for errors throughout the test and assigns a score by subtracting error points from 100.

As noted earlier, the purpose of this study is to provide an estimate of the interrater, sampling and net reliability of the current California road test. These estimates, in turn, are to provide a baseline for assessing whether the new experimental test described in Hagge (1994) represents an improvement. The present study also estimates the difficulty level (e.g., failure rate) of the current test for subsequent comparison with the new test, which was intended to be more stringent and difficult.

METHODS

Research Design

The Research and Development Section performed a preliminary study for the purpose of choosing 6 representative sites from a sample of 30 (Williams & Shumaker, 1994). The thirty field offices were instructed to route to the R&D section copies of all drive test score sheets for the week of 4/29/93.

Mean scores, points off, and fail rates were compared by office and examiner for the 30 sites. The 43 skills on the class C drive test were grouped into 12 sections, such as mechanical operation and equipment use, on the score sheet. Research used these sections as "items" and coded the total points off by adding each of the 12 sections.

The mean fail rate for the entire preliminary sample, including DQs, was .30 and the average number wrong was 35.33. Research selected offices which were close to the average of the 30 offices and which also exhibited similar score profiles on the 12 items. The six chosen sites were: Fullerton, Oxnard, West Covina, Westminster, San Jose, and San Mateo. Additional considerations in office selection were logistical feasibility and the need for both southern and northern regions to be represented. Further details are contained in Williams and Shumaker (1994).

Each DMV field office has a primary and an alternate drive route, and these two routes are used as parallel test routes for computing sampling reliability. The test-retest design counterbalanced the sequence of the two routes and administered the tests contiguous in time (no time lag) between the trials. An overall schematic of the research design is presented in Appendix A. It was determined that a sample of 50 drivers per site was needed in order to have sufficient statistical power. With 6 sites, this required a total sample of 300 drivers. For reasons explained below, that number dropped to 284 for the reliability analyses.

Each office manager chose two examiners for the study and designated one as LRE1 and one as LRE2. Both examiners accompanied the driver with one sitting in the front and one in the back. The front-seat examiner made the licensing decision. The official DMV score was always the higher of the two scores (test and retest) given by the front-seat examiner. The two scores given by the back-seat examiner were used for statistical purposes only.

In contrast to the Stage 3 study (Hagge, 1994), the examiners did not switch seat position after every tested driver. Instead, we created six matrices in which the drive-test assignments and seat positions were randomized within each office. The drive test assignments consisted of either the same examiner on subroute 1 (primary route) and subroute 2 (alternate route) or two different examiners on the same subroute. We randomly assigned the matrices to the study offices.

There were times when the front-seat examiner passed drivers while the back-seat examiner assigned the same driver a failing score. As long as drivers failed on points (losing more than 30 points), this presented no problem for the reliability analysis. A problem arose, however, when the front-seat examiner assigned passing scores to drivers who would have been disqualified by the back-seat examiner. Disqualification's (DQs) occur whenever an error is judged to be so serious as to require termination of the test (e.g. dangerous maneuvers, striking an object, or near accidents). These drivers were not included in the reliability analysis, because the analysis required four completed drive-test score sheets, and DQs do not receive a total test point score. Consequently, this reduced the sample size for the reliability analyses to *N* = 284. DQ treatments were also not included in mean score and points off calculations, because the number of points off at the time of disqualification would not be comparable to number of points lost had the entire test been completed. The DQs, however, were included in the calculation of total test failure rates because this parameter is not affected by the truncation of the test.

Study Particulars

For each driver there were four score sheets and one driver information form. The front-seat examiner took responsibility for collecting the score sheets, filling out the driver information form, and placing the stapled sheets in a study basket after issuing a license. The driver information form was designed to collect demographic data and language fluency information. In offices where a Motor Vehicle Representative issued the license, he or she stapled the five sheets and placed them in the basket. The Drivers License (DL) Managers collected the study sheets on a daily basis and sent them to Research every week.

The DL managers completed a data summary form at the end of the study. This was used for checking that the offices had 50 drivers who had successfully completed a test and retest.

The examiners, DL managers, and field office managers all received a day of training on data collection, information flow, and study procedures. Procedures covered examiner responsibilities, such as scoring independently when working in conjunction with another examiner. DL managers received training in monitoring the daily collection of data. The field office managers were trained in general oversight of study staff and handling the rare occasions when a resistant study subject might be referred to them.

All of the study procedures and reporting requirements were detailed in a protocol, which served as a procedural manual for the field office staff.

Statistical Procedures

Research analyzed the data (total score on the drive test) by calculating three different reliability measures: interrater, interroute (sampling), and net. To determine interrater reliability, we correlated the scores given by both examiners on the same subroute. A correlation of the scores given by the same examiner for the two subroutes produced an interroute reliability. For net reliability, we correlated LRE1's score on one subroute with LRE2's score on the other subroute. In other words, the scores were correlated across examiners and subroutes.

All three reliability measures were computed for the individual offices and for the total sample. We also examined the differences in mean test scores as a function of office, route, and rater. Analysis of variance (ANOVA) and chi square tests were used to test these differences for significance.

The SPSS software package was used for all of the statistical analyses and computations.

RESULTS

Sample Description

The demographic characteristics of the sample are shown in Table 1 below.

Table 1

Current Class C Demographics Percentages

Variable	San Jose	San Mateo	Fullerton	Westminster	West Covina	Oxnard	Total
Gender							
Males	60.6	26.5	56.2	33.3	28.4	49.3	43.4
Females	24.2	44.1	22.9	33.3	45.9	46.4	35.3
Missing	15.2	29.4	21.0	33.3	25.7	4.3	21.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Mean age	24.9	26.5	24.6	24.1	27.2	23.7	25.2
Ethnicity							
Hispanic	59.1	23.5	56.2	25.0	28.4	62.3	42.5
Asian	21.2	36.8	22.9	31.7	28.4	8.7	23.1
Other	19.7	39.7	21.0	43.3	43.2	29.0	34.4
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Foreign license surrendered?							
yes	0.0	2.9	1.0	5.0	0.0	0.0	1.4
no	100.0	97.1	86.7	93.3	100.0	97.1	95.0
missing	0.0	0.0	12.4	1.7	0.0	2.9	3.6
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
English language fluency							
Good	62.1	61.8	48.6	53.3	66.2	59.4	57.9
Marginal	19.7	25.0	2.9	18.3	16.2	11.6	14.5
Poor	18.2	13.2	48.6	28.3	17.6	29.0	27.6
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Knowledge Test language							
English	50.0	83.8	54.3	48.3	63.5	71.0	61.5
Non-English	50.0	16.2	45.7	48.3	36.5	29.0	38.0
Missing	0.0	0.0	0.0	3.3	0.0	0.0	0.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Percentage Passing							
Male	52.5	83.3	62.7	85.0	71.4	61.8	65.6
Female	62.5	63.3	70.8	90.0	76.5	53.1	68.6
Total	55.4	70.8	65.1	87.5	74.5	57.6	67.0

Of the total sample, 43.4 percent were males and 35.3 percent were females. Gender information was missing on the remaining 21.3 percent. Assuming that the missing gender group was comprised of the same proportion of males and females as the preceding, then the estimated proportion of males and females was, respectively, 55% and 45%. Women tended to do slightly better than men on the drive test. Of the females, 68.6 percent passed, while 65.6 percent of the males passed.

The predominant ethnic group was Hispanic (42.5 percent). Asians made up 23.1 percent of the sample and Others 34.4 percent. (For the purposes of the study, Caucasians and African Americans were included in the "Other" group).

While 38 percent of the subjects took the knowledge test in a non-English language, only 1.4 percent of the sample surrendered a foreign license. Examiners were asked to judge language fluency of the drivers along a three-point scale. Of the total sample, 57.9 percent were rated good, 14.5 percent marginal, and 27.6 percent poor in English language fluency.

The office with the highest percentage of Hispanics was Oxnard (62.3 percent), whereas San Mateo had the highest percentage of Asians (36.8 percent). Westminster and West Covina had the highest percentages of Others at 43.3 and 43.2 percent, respectively. The majority of "Others" were presumably Caucasian.

West Covina had the highest percentage (66.2 percent) of drivers with "good" English language fluency, while Fullerton had the highest percentage (48.6 percent) of drivers judged to have "poor" English fluency. San Mateo had the highest percentage (83.8 percent) of drivers taking the knowledge test in English, and San Jose had the highest percentage of drivers taking the non-English test (50 percent).

Oxnard had the youngest drivers (23.7 mean age) and San Mateo had the oldest drivers (26.5 mean age). The mean age for the total sample was 25.2. Fifteen percent of the drivers for the total sample were under 18 years of age.

<u>Fail Rates</u>

<u>Total sample</u>. Various test performance measures for the total sample are summarized in Table 2. The overall fail rate was .38. Test results were pooled to include front- and back-seat scores and test and retest scores. When DQs were excluded, note that the fail rate dropped to .24 and the average score was 76.20. Obviously, DQs contributed heavily to drive-test failures since the fail rate including DQs was substantially higher than the fail rate when they were excluded.

With DQs included, fail rates by field office ranged from .27 for Westminster to .49 for San Jose. Excluding DQs, fail rates varied from a low of .11 for Fullerton to a high of .41 for San Jose.

Table 2

Current Class C Results for Frequencies, Fail Rate, Average Number Wrong, and Mean Score by Field Office

A. DQs Included

Field office	Passes	Fails	Fail rate	N
Total	969	587	.38	1556
San Jose	126	119	.49	245
San Mateo	174	73	.30	247
Fullerton	205	133	.39	338
Westminster	165	61	.27	226
West Covina	171	89	.34	260
Oxnard	128	112	.47	240

B. DQs Excluded

Field office	Passes	Fails	Fail rate	Average # wrong	Mean score	N
Total	955	306	.24	23.80	76.20	1261
San Jose	126	87	.41	29.38	70.62	213
San Mateo	170	41	.19	22.74	77.26	211
Fullerton	201	24	.11	16.77	83.23	225
Westminster	164	42	.20	24.98	75.02	206
West Covina	170	38	.18	21.37	78.63	208
Oxnard	125	74	.37	28.23	71.77	198

Front-seat examiner tests. Fail rates and average number wrong were also calculated for front-seat examiners on the first test. The results are summarized in Table 3. These measures were calculated in order to provide results which more closely approximated the standard road test procedure. In other words the author did not pool all scores across seat position and test/retest status as before. Using this method, the fail rate was .44 for the current drive test. At the field office level, fail rates ranged from a low of .30 for Westminster to a high of .59 for Oxnard. All of the above results included DQs.

When DQs were excluded, the fail rate dropped to .26 for all front-seat, first, test scores. The overall average score was 75.4, with individual office scores ranging from 82.3 (Fullerton) to 68.7 (Oxnard).

Table 3

Current Class C Results for Fail Rate
for Front Seat Examiner on First Test

A. DQs Included

Field office	Fail rate	N
Total	.44	443
San Jose	.49	67
San Mateo	.32	68
Fullerton	.45	105
Westminster	.30	61
West Covina	.45	73
Oxnard	.59	69

B. DQs Excluded

Field office	Fail rate	Average # wrong	Standard deviation	N
Total	.26	24.60	13.39	337
San Jose	.38	30.96	16.55	55
San Mateo	.18	22.46	10.43	56
Fullerton	.11	17.72	10.11	65
Westminster	.20	24.61	10.61	54
West Covina	.27	22.22	10.13	55
Oxnard	.46	31.25	16.14	52

Test Result Frequencies and Percentages

Table 4 shows test results pooled across seat position and test/retest status. The failures for the entire sample are split evenly between DQs and failures on points. Fullerton had the highest percentage of DQs at 83 percent of its total failures. San Jose had the highest percentage of point failures, representing 73 percent of total failures for the office.

Table 4

Current Class C Results for Frequencies and Fail Rate

Field office	Total N	Passes	Discretionary passes	Overall	DQs		Point failures		Total fails
					N	% of total fails	N	% of total fails	
Total	1555	969	15	.38	292	50%	294	50%	586
San Jose	245	126	1	.48	32	27%	87	73%	119
San Mateo	247	174	4	.30	36	49%	37	51%	73
Fullerton	338	205	3	.39	111	83%	22	17%	133
Westminster	226	165	1	.27	20	33%	41	67%	61
West Covina	260	171	1	.34	51	58%	37	42%	88
Oxnard	240	128	5	.47	42	37%	70	63%	112

An explanation is in order for the column of Table 4 labeled "discretionary passes." Discretionary passes occur when examiners give drivers a score of 69, but still pass them. Since only 1.5 percent of the passes in this sample were discretionary, it does not appear to be a widely used practice.

Results by Route

The pooled fail rates and average number wrong are shown in Table 5 by route.

Table 5

Current Class C Results by Route for Fail Rate (Including DQs)

Field office	Route	Fail rate	N
Total	1	.38	762
" "	2	.39	794
San Jose	1	.52	122
" "	2	.45	123
San Mateo	1	.31	123
" "	2	.31	124
Fullerton	1	.24	154
" "	2	.54	184
Westminster	1	.31	115
" "	2	.23	111
West Covina	1	.39	132
" "	2	.30	128
Oxnard	1	.52	116
" "	2	.45	124

These results, which include DQs, indicate that the fail rates for the two routes were almost identical for all offices combined (.38 vs. .39).

Table 6 shows similar comparison of pooled fail rates by route when DQs are excluded. As expected, the fail rate declined for both routes (.26 and .22).

Table 6

Current Class C Results by Route for Fail Rate (Excluding DQs)

Field office	Route	Fail rate	Average # wrong	Standard deviation	N
Total	1	.26	24.36	13.41	642
" "	2	.22	23.22	12.83	619
San Jose	1	.43	29.74	16.68	102
" "	2	.39	29.05	15.24	111
San Mateo	1	.21	22.93	10.01	108
" "	2	.17	22.54	12.25	103
Fullerton	1	.07	16.30	11.62	125
" "	2	.15	17.37	11.05	100
Westminster	1	.22	25.59	10.36	101
" "	2	.19	24.39	11.70	105
West Covina	1	.23	22.90	10.96	104
" "	2	.13	19.85	9.84	104
Oxnard	1	.45	30.63	14.42	102
" "	2	.29	25.69	12.76	96

Reliability Estimates on Test Means and Passes vs. Failure Rates

This section presents the results of the test reliability analysis. Recall that three types of reliability were computed: interrater, interroute and net. The results are summarized in Table 7.

Table 7

Current Class C Reliabilities

A. Interrater Reliability: LRE 1 x LRE 2 by route (total score)

Office	Route 1	Route 2	Average of 2 routes
San Jose	.64	.82	.73
San Mateo	.59	.73	.66
Fullerton	.38	.66	.52
Westminster	.81	.80	.80
West Covina	.67	.65	.66
Oxnard	.65	.42	.54
Total	.67	.72	.69

Table 7 (continued)

B. Interroute Reliability: Route 1 x Route 2 (total score)

Office	LRE 1	LRE 2	Within-LRE average
San Jose	.66	.71	.68
San Mateo	.57	.73	.65
Fullerton	.64	.71	.67
Westminster	.80	.69	.74
West Covina	.73	.58	.66
Oxnard	.53	.68	.60
Total	.66	.67	.66

C. Net Reliability (total score)

Office	LRE 1/route 1 x LRE 2/route 2	LRE 1/route 2 x LRE 2/route 1	Average
San Jose	.59	.74	.67
San Mateo	.47	.41	.44
Fullerton	.50	.43	.47
Westminster	.66	.75	.71
West Covina	.57	.45	.51
Oxnard	.57	.56	.56
Total	.59	.60	.60

Interrater reliability for the total sample was .69, which matches the interrater reliability reported by Ratz (1978) for the then current road test. The interrater reliabilities for the individual field offices ranged from .52 for Fullerton to .80 for Westminster.

Interroute reliability was .66 for the total sample. San Mateo had the lowest interroute reliability at .48, while Westminster had the highest at .74.

Net reliability was .60 for the total sample. San Mateo had the lowest net reliability at .44 and Westminster the highest at .71.

Given the above differences between offices in test reliabilities, it is of interest to determine whether these differences are statistically significant. If they are not, the range of variation is consistent with random sampling from a common population value. If, on the other hand, they are significantly different, then one can conclude that the offices differ in the reliability of their drive tests. The three sets of reliabilities were evaluated using the z-transformation chi-square approach described in Snedecor and Cochran (1976). This technique indicated that the variations between offices for each of the three types of reliability were statistically significant (*p*<.005), as shown in Table 8. It can therefore be concluded that there is some difference in the reliabilities of the road test given by different offices.

Table 8

Significance Tests of Differences in Test Reliability by Office
(*z*-Transformed Reliability Coefficients)

Reliability component	x^2	*df*	*p*
Interrater	18.9	5	.003
Interroute	19.0	5	.003
Net	26.8	5	<.001

Table 9 shows the reliability results using pass vs. fail as the variable. Interrater, interroute, and net reliabilities are all lower for pass vs. failure criterion than for score. It seems likely, however, that these pass vs. fail reliabilities have been attenuated by the exclusion of DQs from the analysis. DQs tend to represent the most aberrant errors and skill deficiencies––factors which should be easiest for LREs to assess and concur on. Another factor contributing to the lower reliabilities is the dichotomous nature of the pass vs. fail scale which tends to be a less sensitive measure than mean scores.

Table 9

Current Class C Correlations Using Pass/Fail as Variable
(Excluding DQs)

	Interrater	Interrroute	Net
Total	.55	.52	.45

Differences Between Offices on Test Scores

Analysis of variance (ANOVA) was used to determine if there were significant differences between offices on mean score on first tests given by front-seat examiners as shown in Table 10.

Table 10

Current Class C ANOVA Results of Differences in Mean Score for Front-Seat Examiner on First Test

Field office	Mean score	SD	*N*
Total	75.44	13.53	337
San Jose	68.85	16.97	55
San Mateo	77.68	10.14	56
Fullerton	82.28	10.11	65
Westminster	75.39	10.61	54
West Covina	77.78	10.13	55
Oxnard	69.02	16.61	52

Score by field office: *F* = 10.12, Sig = <.0001, Eta squared = .1326

The analysis of variance indicated that the variation in mean score among the offices was highly significant (*F* = 10.12; *df* = 5, 331; *p* < .001). It is therefore safe to conclude that the test score difference represented real differences in applicant performance, scoring procedures or both rather than sampling error. This finding is consistent with those of Williams and Shumaker (1994), who also found significant score differences in their analysis of 30 field offices.

A chi square test was used to evaluate diffferences in failure rates between offices on first tests given by front-seat examiners (Table 11). There were no significant differences (*p* > .50) in failure rates on first tests.

Table 11

Fail Rate Differences on First Test for Front-Seat Examiners (Including DQs)

Field office	Range of LRE differences (lowest vs. highest)	Chi square	*df*	*p*	N
San Jose	.14	0.735	1	.391	67
San Mateo	.14	1.102	1	.294	68
Fullerton	.08	0.457	1	.499	105
Westminster	.01	0.000	1	1.000	61
West Covina	.12	0.658	1	.417	73
Oxnard	.01	0.000	1	1.000	69
Total		2.952	5	.710	443

Significance of Differences Between Routes and Raters on Test Score

As stated earlier, each field office is supposed to have a primary and an alternate route which are comparable. Are they in fact the same or are they significantly different in terms of difficulty level? The reader is reminded that this analysis is concerned with differences in the mean test scores and fail rates and not with the similarity in score profiles between routes (which was addressed in the section on reliability). According to the results of the analysis of variance shown in Table 12, there is a suggestive difference between route 1 and 2 (F = 3.75; p = .054, df = 1, 335) but it is very minimal.

Table 12

Current Class C ANOVA Results for
Front-Seat Examiner on First Test

Variable	Mean score	*N*	*F*	Significance of *F*
Route 1	74.15	185	3.75	.054
Route 2	77.01	152		
LRE 1	75.74	173	0.17	.676
LRE 2	75.12	164		

As explained previously, each field office chose one examiner to be LRE1 and one to be LRE2. Although the decision may have been an arbitrary one, there also may have been reasons to assign one examiner to be the first LRE and another to be the second. It appeared that LRE1 was often the lead person.

Table 12 shows the analysis of variance results when front-seat scores are compared for the two groups. There is no significant difference between LRE1 and LRE2 (F = 0.17; p = .676; df = 1, 335). This result is hardly surprising since the mean scores are within one percentage point of each other.

The above analysis does not provide a very sensitive measure of differences between examiners because the aggregation of individual LREs into LRE1 vs. LRE2 is rather arbitrary and ignores all of the individual LRE variance within each group. A more sensitive measure might be to simply compare the scores or fail rates among all of the LREs in the study. This, however, would confound differences between offices and applicants with differences between examiners. In addition, we have already shown that the offices differ significantly in test scores and reliabilities.

Many of these problems can be circumvented by evaluating pass rate differences between LREs within each office. Since there were 6 offices, this resulted in 6 separate chi-squares and *p* values - one for each office. The results are shown in Table 13. Although the results for a given office are limited by the small number of LREs and subjects per office, it is possible to derive a more powerful test of these differences by combining the various chi-squares into a single composite test using the additivity property of the chi-square statistic (Hedges & Olkin, 1985). This produced a highly non-significant result ($p > .50$), providing no evidence that LREs differ in their pass/fail scoring standards.

Table 13

Fail Rate Differences Between LREs Within Office
(Including DQs)

Field office	Range of differences	Chi square	*df*	*p*	N
San Jose	.11	2.547	1	.110	245
San Mateo	.03	0.056	1	.812	247
Fullerton	.03	0.128	1	.721	338
Westminster	.02	0.090	1	.764	226
West Covina	.08	1.708	1	.191	260
Oxnard	.01	0.017	1	.897	240
Total		3.815	5	.650	1556

DISCUSSION

When test results were pooled to include front- and back-seat scores and test and retest scores, the fail rate was .38 for the total sample. When calculations were limited to front-seat examiners on the first test (as is the standard practice) the current test fail rate increased to .44.

Failures were evenly split between point failures and DQs but the mixture differed by office.

The overall mean test score (front seat LRE–first route) for all offices after removing DQs was 75.40.

Analysis of variance results showed that test scores differed significantly by office. This result replicates the finding of Williams and Shumaker's (1994) evaluation of 30 field offices. Although this finding could reflect differences in examiner and office scoring, the likelihood of regional differences in applicant skill precludes an unequivocal interpretation. The analysis of LRE fail rates within each office produced no evidence that LREs differed in pass vs. fail standards.

Current test reliabilities were respectable. Using total score, interrater reliability was .69, interroute reliability was .66, and net reliability was .60. The interrater reliability is remarkably similar to what Ratz found for California DMV passenger tests in the late 1970s (Ratz, 1978). The differences in test reliability across offices were statistically significant, indicating that the offices are not equivalent in terms of test reliability.

The primary purpose of the present study is to provide a normative baseline for evaluating the improved road test for Stage 3 of the project (Hagge, 1994). In evaluating the test failure rates and reliabilities reported here for the current test, the possibility of some artifactual enhancement must be acknowledged. The LREs received a special one-day refresher course and were aware that their offices were part of a special project. The unexpectantly high test failure rates may, at least in part, be attributed to this phenomenon. We know, for example, that the failure rates for these same 6 offices were substantially lower (32%) prior to their being selected for the experimental project.

REFERENCES

California Department of Motor Vehicles. (1990). *Summary of the conference on driver competency assessment*. Washington, DC: Author.

Engel, G. R., Paskaruk, S., & Green, N. (1979). *Driver education evaluation tests*. Ottawa, Ontario: Peat Marnick and Partners.

Forbes, T. W., Nolan, R. O., Schmidt, F. L., & Vanosdall, F. E. (1975). *Driver performance measurement based on dynamic behavior patterns in rural, urban, suburban and freeway traffic*. East Lansing, MI: Department of Psychology and Highway Traffic Safety Center, Continuing Education Service, Michigan State University.

Hagge, R. (1994). *The California driver performance evaluation project: An evaluation of a new driver licensing road test*. Sacramento, CA: Department of Motor Vehicles.

Hedges, L., & Olkin, I. (1984). Nonparametric estimators of effect size in meta-analysis. *Psychological Bulletin*, *96*(3), 573-580.

Jones, M. H. (1978). *Driver performance measures for the safe performance curriculum* (Contract No. DOT-HS-01263). Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration.

McGlade, F. (1963). *Testing driving performance developmental and validation technique*. Research Association, National Commission of Safety Education, and National Education Association.

McKnight, J., & Stewart, M. (1990). *Development of a competency based driver license testing system*. Sacramento, CA: Department of Motor Vehicles.

McPherson, K., & McKnight, A. J. (1981). *Automobile driver on-road performance test: Vol. I final report*. (contract No. DOT-HS-5-0114315-01165). Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration.

Peck, R. (In Press). *Driver licensing and highway safety*. Sacramento, CA: Department of Motor Vehicles.

Ratz, M. (1978) *An evaluation of the California drive test in theme and variation volume II: Final report*. Sacramento, CA: California Department of Motor Vehicles.

Snedecor, G., & Cochran, W. (1976). *Statistical methods* (6th ed.). Ames: Iowa State University Press.

Vanosdall, R. E., Allen, T. M., Pawlowski, J. J., Rohren, J. M., Nolan, R. O., Smith, D. L., Rudisill, M., Specht, P., Hochmuth, M., Spool, M., & Diffley, G. (1977). *Michigan road test evaluation study: Phase III--evaluation study* (final report). MI: Michigan State University Highway Traffic Safety Center and Department of Psychology.

Williams, R., & Shumaker, N. (1994). *Class C drive test baseline study: Preliminary report*. Sacramento, CA: Department of Motor Vehicles.